UCLA ENGINEERING
Henry Samueli School of Engineering and Applied Science
Birthplace of the Internet

Learning to Discriminate in the Wild: Representation-Learning Network for Nuisance-Invariant Image Comparison

Nikolaos Karianakis (UCLA), Yizhou Wang (PKU), Stefano Soatto (UCLA)

UCLA Computer Vision Lab, Peking University Computer Vision and Digital Art Research Group

Overview

- The problem of *nuisance variability* is very acute in Computer Vision, where even the same object can yield a large variety of images depending on many factors, such as vantage point, lightning conditions and partial occlusion.
- Many representation and deep learning architectures have shown the ability to learn the intrinsic variability despite the presence of significant nuisance variability.
- We are interested in testing the hypothesis that a representation learning architecture is able to train away nuisance variability, which is present in images, owing to changes of viewpoint and illumination, noise, and miscellaneous defects.
- We design an architecture based on the Gated Restricted Boltzmann Machine and we challenge it in Computer Vision problems such as:
  - Occlusion Detection
  - Image Segmentation

Training away transformations

- Our generic one-layer learning architecture is based on the Gated Restricted Boltzmann machine. We use a factorized "similarity" measure among two binary images x, y and the model's hidden layer h. The joint probability distribution has energy function:

$$E(\mathbf{x},\mathbf{y},\mathbf{h};\theta) = -\sum_{f=1}^{F}\left(\sum_{i=1}^{I} u_{if}x_i\right)\left(\sum_{j=1}^{J} v_{jf}y_j\right)\left(\sum_{k=1}^{K} w_{kf}h_k\right) - \sum_{i=1}^{I} a_i x_i - \sum_{j=1}^{J} b_j y_j - \sum_{k=1}^{K} c_k h_k,$$

  where ϑ = {U, V, W, a, b, c} are the model parameters.

- Training with *3-way Contrastive Divergence* provides a mechanism that recognize and is able to eliminate small affine transformations and lightning changes between images. The joint log-likelihood is:

$$log\ p(\mathbf{x},\mathbf{y}) = -logZ + \sum_{i=1}^{I} a_i x_i + \sum_{j=1}^{J} b_j y_j + \sum_{k=1}^{K} log\left(1 + e^{c_k + \sum_{f=1}^{F} w_{kf}\left(\sum_{i=1}^{I} u_{if}x_i\right)\left(\sum_{j=1}^{J} v_{jf}y_j\right)}\right)$$

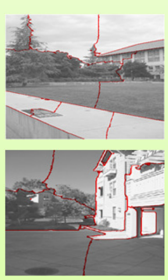
- A semi-metric is used as a nuisance-invariant distance between two images:

$$d(\mathbf{x},\mathbf{y}) = -log\ p(\mathbf{x},\mathbf{y}) - log\ p(\mathbf{y},\mathbf{x}) + log\ p(\mathbf{x},\mathbf{x}) + log\ p(\mathbf{y},\mathbf{y})$$

Semantic Image Segmentation

- The Gated RBM is trained with pairs of random images related with affine transformations, different illumination and scale. This way, the model can recognize patches coming from the same object.
- Comparisons of the neighboring patches in a single image reveal the "semantic" boundaries, that is separating curves between *different* objects in the scene. Not meaningful edges, such as the tiles and the texture within the clouds and the grass, are successfully disregarded by our algorithm.
- The Gated RBM's "similarity" map among all the patches that cover the image drives Normalized Cuts and combined yield a semantic segmentation.

Semantic boundary detection



Normalized-cuts

Gated RBM-driven Normalized-cuts segmentation

Occlusion Detection

- Occlusion detection is the binary classification task of determining the *co-visibility* from different images (e.g. two sequential video frames) of the same scene.
- The model is trained with training image pairs related by affine transformations, shifts and rotations, scale and illumination variation. The first factors intend to deal with different vantage points where these images are captured from, while the latter one with different lightning conditions (background variation subtraction).
- The trained model is tested on sequential video frames from Middlebury, Berkeley MOSEG and UCL Optical Flow datasets. Comparisons between corresponding patches in the two frames yield the occluded areas, after thresholding the semi-metric d.

Baseline algorithm based on differences of patchwise intensity averages.

Gated RBM trained on shifts.

Gated RBM trained on shifts and rotations.

Gated RBM trained on shifts, rotations, affine, scale and illumination variation.

Gated RBM trained on shifts, rotations, affine, scale and illumination variation, having considered many superpixel maps.

- The superpixels capture uniform intensity areas in different scales. Our conjecture is that with high probability their pixels back-project to points in the scene with the same motion.
- Extracting multiple joint superpixels improves the performance.